ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Machine Learning Methods and Data Mining in Computer Security

S.Kavitha

Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation Green Fileds, Vaddeswaram, A.P. – 522302.

Abstract.

Data Mining is considered as an interdisciplinary discipline. Individuals attempt to address significant scale of DM problems using Artificial Intelligence (AI) techniques for real-world applications such as industrial, social, healthcare, predictive toxicology, and bio, chemo-informatics. This paper intends to address some confusion related to overlapping notions in Data Mining techniques and proposes a survey on one AI technique(Decision Trees), to solve specific Data Mining problems,and also a work on Data Mining for security applications. Intrusion Detection (ID) is the main research area in network security. It involves the monitoring of the events occurring in a computer system and its network. Data mining a new technology applied to ID to invent a new , attching from the massive network data as well as to deduct the strain of the manual compilations of the intrusion and normal behaviour patterns.

Keywords: Data Mining, DecisionTrees, Intrusion Detection, Classification, Clustering.Artificial Intelligence,

1. Introduction

Artificial Intelligence techniques show a new and up-rising development under the new evolution of storage and processing performances of computing machines. Moreover, the data stored starts to signal a new crisis, not only the availability, but also quality, sources, and mainly the meaning are new issues the users are interested in. Noisy, untrusting, heterogeneous data cover and hide the golden nuggets of the patterns and the predictive abilities every user shall be looking to. This paper explains key works in Mining to identify common areas of overlap, applications, and potential sources of confusion arising from

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

different perspective. A case study focused on dynamics of main techniques and applications, of machine learning.

2. General Concepts

We introduce in this most used terms in Artificial Intelligence, Knowledge Discovery , as defined in the majority literature sources.

Information is generally defined [1, 2] as Data transformed and presented in a human-readable format with the intent of uncovering insights, such as patterns or rules

Data is defined [1] as the Facts regarding things such as people, objects, events, which can be described as the other face of the information in its numerical form, which can be digitally transmitted or processed.

Models are defined [3] as creating a representation of patterns worthy of being standard ones.

Knowledge is the theoretical and practical comprehension [2] of a certain domain that supports making decisions.

Intelligence refers to the capacity to acquire knowledge, comprehend, and devise solutions for issues within a particular domain

Artificial Intelligence dealing with helping computing machines to provide solutions for dilemmas facing scientists and building intelligent systems in a way that human acts.

2.1 Machine Learning

Techniques refer to a computer's capacity to enhance its performance through past outcomes observed as contemporary research trends. These methods are extensively applied to real-world challenges. Figure 1 outlines a selection of Machine Learning techniques [4] employed in data mining

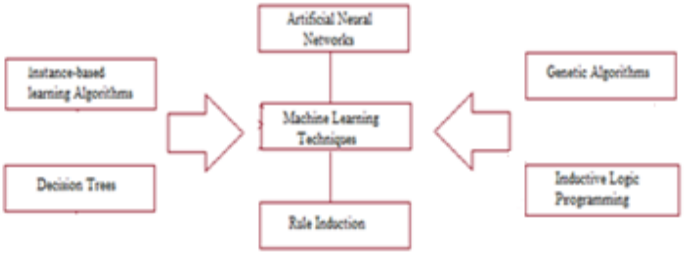

fig 1.Machine Learning Techniques Applied to Data Mining

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

2.2 Why Data Mining?

The volume of data to be growing at an exponential rate in most domains associated with information processing, necessitating the of techniques to extract and acquire knowledge from databases is still crucial.

Data Mining (DM) defines the automated extraction from databases [5]. DM problems addressed by intelligent systems are:

1. **Diagnosis**: Defining malfunctions based on an object behaviour then recommending solutions.
2. **Pattern Recognition**: Identification of objects and images by their shapes, forms, outlines, colour, surface, texture, temperature, or other attribute, usually by automatic means.
3. **Prediction**: Predicting the behaviour of an object future pertaining to its past one.
4. **Classification**: Assigning an object to a pre-defined class.
5. **Clustering**: Dividing a heterogeneous group of objects into homogeneous subgroups.
6. **Optimisation**: Improvising solutions until an optimal one is found.
7. **Control**: Governing of object to meet specified requirements in the real-time [5].

2.3 ML Techniques to DM Tasks

Concerning their suitability for data mining tasks, can be categorized according to their suitability, performance, and impact.

1. **Neural Networks for DM**: Artificial Neural Networks (ANN) find extensive applications in various fields, including pathology, biology,imageprocessing,numerical analysis, and control systems.
2. **Genetic Algorithms for DM**: Genetic Algorithms, of evolutionary computation techniques, are widely recognized as effective problem-solving methods in the domains of chemistry, biotechnology, movement prediction, bioinformatics, and adaptive control for operational systems
3. **Inductive Logic Programming for DM:**Inductive Logic Programming (ILP) is typically applied to a narrower range of applications . Nonetheless, it has found use in various areas, including disease diagnosis, classification, clustering, and the control of robotics systems
4. **Rule Induction for DM**: Symbolic rule induction exhibits a degree of applicability in optimization tasks, with Semantic query optimization serving as a notable example

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

5. **Decision Trees for DM**: Decision Trees (DT) serve as potent data mining tools for addressing real-world challenges, including prediction and classification tasks. Furthermore, induction process lead to the derivation of control rules
6. **Instance-based Algorithms for DM**: Instance-Based Learning (IBL) is the method of generalizing a new instance to be classified by referencing the stored training examples, a method commonly applied in classification tasks

3. Classification in Data Mining

The classification algorithms in data mining has garnered substantial attention. Comparing classification algorithms is a ongoing challenge for several reasons. The performance can defined as various ways, including accuracy, cost, and reliability. A consistent methodology must be chosen for comparing the measured parameters.

Choosing the best algorithm for a given dataset is a frequently encountered and widely recognized challenge. This procedure requires a series of methodological decisions. In this study, the primary emphasis is on decision tree a within the realm of classification methods, used to measure the performance of classification. Classification techniques can be classified into five distinct groups, each based on different mathematical concepts. These categories include statistical-based, distance-based, decision tree-based, neural network-based, and rule-based methods."

3.1 Evaluating the Effectiveness of Decision Tree: Decision Trees are among the most commonly utilized supervised classification techniques. The process of learning and classifying through decision tree induction is recognized for its simplicity and speed, making it applicable across various domains. Decision trees essentially operate as hierarchical structures that classify instances based on their feature values. Each node of decision tree corresponds to a feature to be classified, and each branch has a potential value for that feature[2]. Classification begins at the root node, with instances sorted based on their feature values..The algorithm,summarized as follows.

1. Create node N;
2. If all samples fall under to the same class, denoted as C, then
3. Return N as a leaf node, marked with the class label C;
4. If attribute-list is empty then

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

5. Return N as a leaf node, marked with the class label C

6. Select test-attribute, the attribute among attribute-list

with the moreinformation gain;

7. Label node N with test-attribute;

8. For each known value ai of test-attribute

9. Grow a branch from node N for the condition

test-attribute= ai;

11. If si is empty, then attach a leaf labeled with the most frequent class in the samples.

12. Else the node by

Generate_decision_tree(si,attribute-list_test-attribute)

Fig.2 Decision tree algorithm.

A Classification Task Case Study

TABLE 1. Weather Data for play tennis

Id	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rain	Mild	High	False	Yes
5	Rain	Cool	Normal	False	Yes
6	Rain	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rain	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	yes
13	Overcast	Hot	Normal	False	Yes
14	Rain	Mild	High	True	No

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

TABLE2. Weather data

Variable	Attribute type	Possible values
Outlook	categorical	{sunny,overcast,rain}
Temparature	categorical	{hot,mild,cool}
Humidity	categorical	{high.normal}
Windy	Binary	{true.false}
Result	----------	{play or not play}

3.2 Algorithms

Decision Tree is a highly valuable and renowned classification method. It offers the advantage of a straightforward process for creating and presenting results [1]. When provided with a dataset containing attributes and their corresponding classes, a decision tree generates sequences of rules that aid in class recognition for decision-making purposes. The most common algorithms for decision trees are ID3, C4.5, and CART. This study reveals that the CART algorithm outperforms the ID3 algorithms in terms of accuracy. The key strength of the CART algorithm lies in its exhaustive evaluation of all potential attribute splits. Once the best split is identified,. Consequently, the CART algorithm has proven to be instrumental in enhancing data classification accuracy

A. ID3

ID3, designed by Ross Quinlan [6], is a straightforward decision tree learning algorithm. It operates on the fundamental principle of constructing the decision tree using a greedy search through the provided datasets.. The algorithm evaluates the attribute's worth using a statistical measure known as information gain.

a) Impurity

When provided with a data table comprising attributes and their associated classes, we can assess the uniformity or diversity [6] of the table concerning the classes. A table is considered pure or homogeneous when it contains only one class. Conversely, if a data table encompasses multiple classes, it is described as impure or heterogeneous. To quantify the level of impurity or entropy,

$$\textbf{Entropy} = \sum -P_j \log_2 P_j \qquad (1)$$

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

The entropy of a completely homogeneous table (consisting of a single class) is zero, as the probability is 1, and log(1) equals 0.

In order to prove the most suitable attribute for a specific node in the tree, information gain is employed. Information gain, denoted as Gain(S, A), measures the attribute A's contribution relative to the set in S

Gain(S,A)= Entropy(s) - $\sum_{v\in values(A)} \frac{|Sv|}{|S|}$ **Entropy(S_v)(2)**

TABLE 3:Information Gain Values of Weather Dataset

Gain	Values
Gain(S,Outlook)	0.247
Gain(S,Temperature)	0.029
Gain(S,Humidity)	0.152
Gain(S,Windy)	0.048

Table 3, Outlook has the more gain, it is used as the root shown in Fig.3

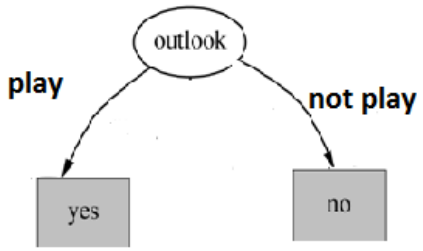

Figure 3. A root node given Weather data

c) procedure to design a decision tree Using computation results, the attribute 'Outlook' is selected to grow the tree. Subsequently, the 'Outlook' attribute is removed from the samples within these sub-nodes, and the Entropy and Gain are calculated to determine the attribute with the more gain value for further tree expansion. Continue this process until the Entropy of the node value zero. When this occurs, the node can no longer be extended as the samples within that node exclusively belong to a same class.

B. C4.5

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

The C4.5 algorithm [7] is an evolution of ID3 and employs the gain ratio as a splitting criterion for data set partitioning. This algorithm introduces a form of normalization to the information gain by incorporating a 'split information' value

a) Measuring Splitting Criteria

To identify most suitable attribute for specific node within the Information Gain is employed. Information Gain, denoted as Gain(S, A) for an attribute A in relation to a set of examples S, is as

$$\textbf{Gain(S,A)= Entropy(s) - } \sum_{v \in values(A)} \frac{|Sv|}{|S|} \textbf{Entropy}(S_v) \textbf{(3)}$$

Where Values(A) represents the set of all values for attribute A, and Sv denotes the subset of S in which A holds the value v (i.e., Sv = { A(s) = v}). The first term in the Gain equation corresponds to the initial collection S, while the second term denotes the expected entropy value after partitioning S with attribute A. The expected entropy, as described in the second term, is essentially the summation of entropies for each subset, eachthe fraction of examples |Sv| / |S|.n

$$\textbf{Split Information(S,A)= } -\sum_{i=1} \frac{|St|}{|S|} \log_2 \frac{|St|}{|S|}$$

and

$$\textbf{Gain Ratio(S,A)=} \frac{\textbf{Gain Ratio(S,A)}}{\textbf{Split Information(S,A)}} \textbf{(4)}$$

The process involves selecting a fresh attribute and partitioning the training examples, and this is repeated for every non-terminal descendant node. Attributes previously utilized higher up in the tree are omitted, ensuring that each attribute appears at most once along any branch. This iterative procedure persists for each leaf node until one of two conditions is satisfied

Either every attribute has already been employed along the path in the tree, or the training examples linked to this leaf node all showcase the same target attribute value.(i.e., they have zero entropy).

For attribute selection, the Gain Ratio can be employed. Before computing the Gain Ratio, Split Information must be calculated, as demonstrated in Table 4

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

TABLE 4: Split Information

Split Information	Values
split(S,Outlook)	1.577
split(S,Temperature)	1.362
split(S,Humidity)	1.000
split(S,Windy)	0.985

The Split Information, as displayed in TABLE 4, indicates that 'Outlook' possesses the highest gain ratio, making it the suitable choice for the root node. The decision tree created for both ID3 and C4.5 exhibits a structure resembling that depicted in Fig. 3. Subsequently, the Information Gain is computed, as detailed in TABLE 5.

TABLE5: Gain

Split Information	Values
GainRatio(S,Outlook)	0.156
GainRatio(S,Temperature)	0.021
GainRatio(S,Humidity)	0.152
GainRatio(S,Windy)	0.049

The Gain Ratio, as presented in TABLE 5, reveals that 'Outlook' possesses the highest gain ratio, leading to its selection as the root node, as depicted in Fig. 3. Once the algorithm identifies the optimal attribute, it proceeds to partition the data table based on this attribute. In our sample data, C4.5 split the data table based on the values of 'Outlook

b) Procedure to build tree

Start with the initial samples as the root of the decision tree. Following the calculation, the 'Outlook' attribute is chosen to expand the tree. Subsequently, the 'Outlook' attribute is removed from the samples in these sub-nodes, and split information is computed to further divide the tree using the attribute with the highest gain ratio. This process repeats for all data is classified accurately or until no more attributes are available. Continue this process until the Entropy of the node reaches zero. At this point, the node can no longer be expanded because the samples within it all belong to the same set.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

C. Decision Tree using CART

a) CART [2], an abbreviation for Regression Trees, was introduced by Brieman. It also draws upon Hunt's algorithm. CART is capable of handling both categorical and continuous attributes to construct a decision tree, and it can accommodate missing values[10]. This algorithm utilizes the Gini Index as selection measure for tree construction. CART generates binary splits. The Gini Index does not rely on assumptions like ID3.

b) Impurity To measure degree of Gini Index defined as

$$\textbf{Gini(T) = 1-}\sum^{n}_{j=1}\textbf{P}_j^2\textbf{(5)}$$

Gini Index of a completely homogeneous table containing a single class. Similar to Entropy, the Gini Index also reaches its high value when all classes in the table has equal probabilities. To determine the information gain for A in relation to S, the first is to compute the Gini Index of S.

TABLE 6. Gain Values for CART

Gain	**Values**
Gain(S,Outlook)	0.116
Gain(S,Temperature)	-0.004
Gain(S,Humidity)	0.091
Gain(S,Windy)	0.030

4. RESULTS

Comparing classification algorithms can be challenging because the effectiveness of an algorithm can vary depending on the dataset. Performance evaluation for classification algorithms typically revolves around accuracy, although other factors such as computational time and space are also considered to gain a comprehensive understanding of each algorithm.

The figure below illustrates the final tree for our weather dataset. This tree is designed to handle every attribute values present in training data, each of which is assigned a unique class label. Additional conditions are needed to address the following cases:

Certain child nodes formed in the second stage of the diagram might be vacant, signifying a lack of records associated with these nodes. In such instances, the node is identified as a leaf

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

node and inherits the same class label as the majority class of training records linked to its parent node..

In the second stage, if all the records associated with Dt possess identical attribute values, further splitting is not possible. Under these circumstances, the node is designated as a leaf node, inheriting the label of the majority

class among the training records linked to its parent node..

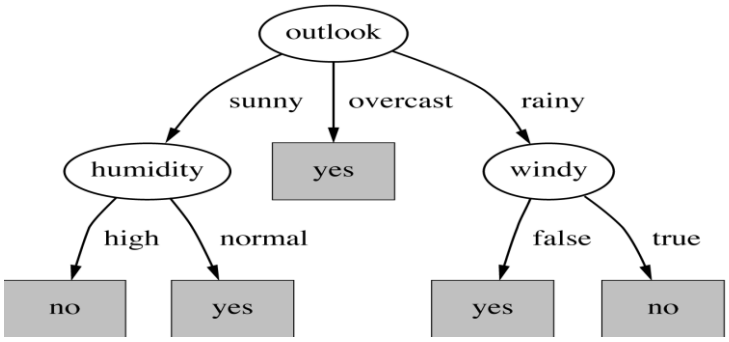

Figure 4. generated decision Tree.

5.Data Mining methods in Intruder Detection

Data mining is is used to uncovering patterns within datasets. It has become a vital for modern businesses as it enables the conversion of data into valuable business intelligence, providing a competitive advantage. Data mining finds applications in various profiling practices, including marketing, surveillance, fraud detection, and scientific discovery. One of its primary purposes is to facilitate the analysis of large collections of behavioral observations.

Data mining technology offers several key advantages:

It can efficiently handle vast amounts of information.

It can reveal concealed and overlooked information.

Data mining involves four primary categories of tasks:

Clustering: This task involves identifying groups and patterns in the data that exhibit similarity, without relying on pre-defined structures.

Classification: It aims to generalize known patterns and apply them to new data.

Regression: This task seeks to find a function that models with minimal error.

Association rule learning: It searches for relationships and associations between variables

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

5.1. Intrusion And Intrusion Detection

Intrusion, in simple words, is an illegal act of entering,seizing, or taking possession of another's property (in this case the property being the computer system It refers to a code that disrupts the smooth flow of traffic on the network or pilfers information from the traffic. The different categories of intrusions can be outlined as follows::-

DoS Attack - A denial-of-service attack (DoS attack) or distributed denial-of-service attack (DDoS attack)[8]to make a computer resource unavailable to its intended users.It generally consists of the concerted efforts of a person or people to prevent an

Internet site or service from functioning efficiently or at all,

temporarily or indefinitely.

Remote to User (R2L) This type of attack involves unauthorized way to access from a remote system to the super user (root) account of the target system. It falls under the category of attacks an attacker sends packets over a network and then exploits vulnerabilities in the machine to gain unauthorized access .

User to Root (U2R) – User to root attack defines the unauthorized access to super user . These exploits are classes of attacks which an attacker starts out with access to a normal user account on the host system and is able to exploit vulnerability to gain root access to the system.

Probing – Probing is a type of attack in which an assailant scans a network to collect information or identify vulnerabilities. An attacker armed with a map of networked machines and their available services can utilize this information to search for potential exploits.

Anomaly Detection: This method involves identifying patterns in a given dataset that deviate from established normal behaviour. These detected patterns, known as anomalies, often provide valuable and actionable insights across various application domains. Anomalies are also known to as outliers, surprises, aberrations, deviations, peculiarities, and more. It entails storing features of a user's typical behaviour in a database and then comparing the user's current behaviour with the stored information in the database[11]

2. **Misuse Detection -** In misuse detection approach, we define abnormal system behaviour at first, and then define any other behaviour, as normal behaviour. It assumes that abnormal behaviour and activity has a simple to define model. Its advantage is simplicity of adding known attacks to the model. Its disadvantage is its inability to recognize unknown attacks.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

Misuse Detection refers to confirming attack incidents by matching features through the attacking feature

library.

5.2 Classification Techniques

The various classification techniques in this study are as follows:-

1. Decision trees: Of a given data sample through various levels of decisions to help reach a final decision [7]

2. Support Machines: SVM first maps the input vector into a higher dimensional feature space and then obtain the optimal separating hyper-plane in the higher dimensional feature space. an SVM classifier is designed for binary classification.

3. Fuzzy logic: It processes the input data from the network and describes measures that are significant to the anomaly detection [8]. Fuzzy logic (or fuzzy set theory) is based on the concept of the fuzzy phenomenon to occur frequently in real world. Fuzzy set theory considers the set membership values for reasoning and the values range between 0 and 1. That is, in fuzzy logic the degree of truth of

a statement can range between 0 and 1 and it is not constrained to the two truth values (i.e. true, false) [10].

4. Naïve Bayes: Naïve Bayesian Networks (NB). The model provides an answer to questions like ‘‘What is the probability that it is a certain type of attack, given some observed system events?”by using conditional probability formula. The structure of a NB is typically represented by a directed acyclic graph(DAG),[10In this context, each node signifies a system variable, and each link conveys the impact of one node on another.. Thus, if there is a link from node A to node B,A directly influences B.

5.3 Clustering Techniques

Clustering basically means that we have to make the group(clusters) from our data so that we can easily find our required data. Clustering is an best way to find hidden methods in data that humans might otherwise miss. Clustering provides some significant advantages.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

6. CONCLUSION

This study explores the usability and performance of machine learning (ML) techniques for solving data mining (DM) problems. The research compares effective of the ID3, C4.5, and CART algorithms. The investigation uses the 'Play Tennis' dataset t of qualitative data when employing decision tree The objective of integrating classification results from various algorithms is to enhance the certainty, precision, and accuracy of system outcomes. Multiple methods have been proposed for creating classifier ensembles. However, despite the numerous approaches, there is still no definitive consensus on the best Classification techniques excel in modeling interactions and are instrumental in the field of network intrusion detection, where data mining plays a pivotal role. The goal is to improve the accuracy of intrusion detection while minimizing false negatives. As data mining is an evolving field, further study and research are warranted.

References

[1]. J. Han, M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[2]. D.Howe. The Free On-line Dictionary of Computing. (1993-2004). Supported by Imperial College Department of Computing, Copyright @1993 by Denis Howe.

[3]. T.Mitchell. Machine Learning. MIT Press and McGraw-Hill, 1997.

[4]. M. A. Bramer. Knowledge Discovery and Data Mining. London: The Institution of Electrical Engineers, 1999.

[5].Jiawei Han and Micheline Kamber, "*Data Mining: Concepts and Techniques*", 2nd ed., Morgan Kaufmann Publishers, 2006.

[6]. M. El-Halees, "Mining Student Data to Analyze Learning Behavior: A Case Study". In Proceedings of the 2008 *International Arab Conference of Information Technology (ACIT2008),* University of Sfax, Tunisia, Dec 15- 18

[7]. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.Uthurusamy, "*Advances in Knowledge Discovery and Data Mining*", AAAI/MIT Press, 1996.

[8]. Alexander D. Korzyk. A Forecasting Model For Internet Security Attacks.

[9]. Simon Hansman and Ray Hunt (2004). A Taxonomy of Network andComputer Attacks

[10].Mrityunjaya Panda and Manas Ranjan Patra. A Comparative Study ofData Mining Algorithms for Intrusion Detection.